DOCUMENT RESUME

ED 161 914 TM 007 606

AUTHOR Long, John; And Others
TITLE An Empirical Investigation of the ESEA Title I Evaluation Systems' Proposed Variance Estimation Procedures for Use With Criterion Referenced Tests.
PUB DATE Mar 78
NOTE 10p.; Paper presented at the Annual Meeting of the American Educational Research Association (62nd, Toronto, Ontario, Canada, March 27-31, 1978)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Compensatory Education Programs; *Criterion Referenced Tests; *Educational Assessment; Elementary Education; *Evaluation Methods; Goodness of Fit; Norm Referenced Tests; *Norms; *Program Evaluation; Reading Achievement; Research Methodology; Statistical Analysis; *Test Interpretation
IDENTIFIERS Elementary Secondary Education Act Title I; *RMC Models

ABSTRACT

An experiment was performed to evaluate the tenability of the assumption in the Elementary Secondary Education Act (ESEA) Title I proposed variance estimation procedures for criterion referenced tests. The assumption is that the ratio of the local to the national standard deviation for the national sample will be the same for the normed test as for the non-normed test. This assumption and the relevant procedure have been proposed only for the study of treatment effects where they depend on the use of aggregated data and not for evaluation at the local level. Two nationally normed reading achievement tests were used, and comparisons were made of actual standard deviations for norm groups with estimates of norm group standard deviations using the ESEA procedure separately for five different grade levels. All five estimates were above the actual values in error by eight percent to 57%. Until further study is made of the errors made in estimating population standard deviations, it appears that the use of non-normed tests in Title I evaluations is a questionable practice. (Author/CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Session 32.11

ED161914

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

An Empirical Investigation of the ESEA Title I Evaluation Systems' Proposed Variance Estimation Procedures For Use With Criterion Referenced Tests

John Long, University of Rhode Island

Stephen Horwitz, University of Rhode Island

Pasquale DeVito, Rhode Island State Dept. of Education

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

John Long
Univ. of Rhode Island

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

Paper presented at the annual meeting of
the American Educational Research Association,
Toronto, March, 1978

TM007 606

Background

Title I of the Elementary Secondary Education Act (ESEA) authorizes financial assistance to local education agencies (LEA's) with concentrations of economically disadvantaged students. These funds are designated specifically for programs designed to meet the special needs of the educationally disadvantaged. Title I legislation requires, in the Education Ammendments of 1974 (P.L. 93-380, sections B, D and F), that the Commissioner of Education provide for the development of standards and evaluation models which yield data that can be aggregated to provide an overall assessment of Title I academic impact. In compliance with these guidelines, the United States Office of Education (USOE), under competative bidding procedures, awarded RMC Research Corporation two contracts for development and refinement of a series of models for evaluating the cognitive outcomes of Title I academic programs. The evaluation models which were developed by RMC have formed the basis for a new "evaluation and reporting system" for ESEA Title I programs. The new evaluation systems are currently being disseminated to school systems around the country by regional Technical Assistance Centers, and it is anticipated that at some future date the use of these evaluation systems will be mandated by USOE for the evaluation of Title I reading and mathematics programs in grades 2 through 12.

One of the unique characteristics of the new evaluation models is that they allow for the use of criterion referenced (or, as they are called in the evaluation models, non-normed) tests in the evaluation of Title I programs. Due to the requirement that all evaluation data be in a form which can be aggregated across schools and districts, in order to use criterion referenced tests for evaluation purposes under the Control group and Regression models a school district would have to have

administered a nationally normed test at some time during the program year. The sole purpose for administering the normed test is so that its standard deviation can be used to estimate the population standard deviation of the criterion referenced test--a key element in the calculation of the programs "treatment effect."

According to the User's Guide to the new evaluation systems, the standard deviation of the national sample for the criterion referenced test is estimated from scores on the nationally normed test

> "under the assumption that the ratio of the local to the national standard deviation will be the same for normed test as for the non-normed test. In other words

$$\frac{S_n}{\sigma_n} = \frac{S_{nn}}{\hat{\sigma}_{nn}}$$

where

S_n = standard deviation of the treatment group scores on the normed test

S_{nn} = standard deviation of the treatment group scores on the non-normed test

σ_n = standard deviation of the norm group scores on the normed test

$\hat{\sigma}_{nn}$ = estimated standard deviation of a nationally representative sample on the non-normed test

Solving for $\hat{\sigma}_{nn}$ the equation becomes

$$\hat{\sigma}_{nn} = \frac{\sigma_n S_{nn}}{S_n} \text{ ."}$$

(User's Guide, 12/76, p. 17.)

The importance of the accuracy of the variance estimation formula can be seen when examining the formula for the program's "treatment effect," the basic measure used in aggregating measures of various programs' effect on cognitive functioning.

$$T.E. = 21.06 \left[\frac{\bar{Y}_t - \hat{\bar{Y}}_t}{\hat{\sigma}_{nn}}\right] \quad (\underline{\text{User's Guide}}, \text{ p. 17.})$$

where

T.E. = Treatment Effect

$\bar{Y}_t$ = Observed mean posttest score of treatment group pupils

$\hat{\bar{Y}}_t$ = Expected mean posttest scores of Treatment group pupils under non-normed test

$\hat{\sigma}_{nn}$ = Estimated standard deviation of a nationally representative sample on the non-normed test

Since a large number of local school districts have both norm and criterion referenced testing programs, the tenability of this assumption, particularly for an atypical population such as those in Title I programs, becomes critical if results of the criterion referenced testing are to be integrated in the Title I evaluation system and aggregated with data collected using norm referenced tests. It should, however, be clearly understood that the credibility of the assumption underlying the variance estimation procedure is not a factor in the use of criterion referenced tests for program evaluation at the local level. It is only at the point where these data are being entered into the uniform evaluation and reporting system, and aggregated across communities, that the assumption becomes critical.

Method and Data Source:

To examine the above assumption directly it would be necessary to administer a norm and a criterion-referenced test to a local Title I group, to administer the same criterion-referenced test to a representative national sample, and then to compare the actually obtained standard deviation of the nationally representative sample on the criterion-referenced test with the estimated standard deviation which would be obtained following the procedures outlined in the User's Guide. In the absence of such a direct test of this assumption, however, a useful estimate of the tenebility of the assumption can be gained by examining the degree to which the ratio of local to national standard deviations on two norm-referenced tests are comparable.

The sample for this study consisted of 318 students who were participating in a Title I reading program in 1976-77 at grades 2, 3, 4, 7 and 8 in a moderate sized Rhode Island community. All of the students were tested with two nationally normed reading tests--the California Achievement Test (CAT) and the Comprehensive Test of Basic Skills (CTBS). Administration of both tests took place within a 2-week period in April, 1976. Although both of these tests have nationally developed norms, for the purposes of this study the CTBS was considered to be a normed test and the CAT was considered to be a criterion-referenced test. The standard deviations for the Total Reading score for the groups tested were calculated and the formula presented previously was then used to "estimate" the population standard deviation for the CAT. Estimated CAT standard deviations were then

compared to the actual CAT standard deviations obtained by inspection from the CAT Bulletin of Technical Data, Number 1.

Results:

Table 1

	S_n	S_{nn}	σ_n	σ_{nn}	$\hat{\sigma}_{nn}$
Grade 2 N=62	9.82	12.86	18.44	19.52	24.14
Grade 3 N=71	10.45	13.29	19.95	16.18	25.37
Grade 4 N=68	12.30	11.02	19.29	16.03	17.28
Grade 7 N=60	6.17	6.82	18.71	16.19	20.68
Grade 8 N=57	10.06	9.65	19.24	16.07	18.46

where

S_n = standard deviation of the Title I students on the normed test (CTBS)

S_{nn} = standard deviation of the Title I students on the "criterian referenced" test (CAT)

σ_n = standard deviation of the norm group on the normed test (CTBS) as obtained from the CTBS Technical Manual

σ_{nn} = standard deviation of the norm group on the "criterion-referenced" test (CAT) as obtained from the CAT Technical Manual

$\hat{\sigma}_{nn}$ = estimated standard deviation of the norm group on the "criterion-referenced" test (CAT) as obtained from the application of the estimation procedure outlined in the User's Guide.

Inspection of Table 1 shows that, at all grade levels studied, proceeding under the assumption that the ratio of local to national standard

deviations would be the same (e.g. that $\frac{S_n}{\sigma_n} = \frac{S_{nn}}{\hat{\sigma}_{nn}}$) led to an estimated standard deviation of the national norm group which was higher than the actual standard deviation for the national norm group obtained by inspection from the publishers Technical Manual. At some grade levels the estimated standard deviation was quite close to the actual standard deviation (e.g. at grade 5 the overestimate was only by 1.25 points, or 8%). At some other grade levels, however, the estimated standard deviations were substantially above the actual standard deviations (e.g. at grade 3 the overestimate was by 9.19 points, or 57%).

Discussion:

The study empirically investigated an assumption of the proposed ESEA Title I evaluation system which allows communities to utilize criterion-referenced (non-normed) tests in the evaluation of their local Title I programs. The assumption--that the population standard deviation of a non-normed test can be accurately estimated since "the ratio of the local to the national standard deviation will be the same for the normed test as for the non-normed test"--is of critical importance since it allows local communities to follow a criterion-referenced testing procedure, to estimate a population standard deviation for their criterion-referenced test, and then to calculate a program's "Treatment Effect" in such a way that the results of their evaluation can then be aggregated. For the student test scores examined in this study, however, estimated and actual national standard deviations were often quite different, estimated standard deviations being systematically larger than actual standard deviations. Since the treatment effect of a Title I program is calculated using the formula,

$$T.E. = 21.06 \left[\frac{\bar{Y}_t - \hat{Y}_t}{\hat{\sigma}_{nn}}\right]$$

the fact that in this case the population standard deviations ($\hat{\sigma}_{nn}$) were systematically overestimated would tend to systematically understate the effects of the Title I program.

It should be noted that had the CAT been chosen as the norm-referenced test and the CTBS chosen as the criterion-referenced (non-normed) test, and the same procedures followed, the converse would have been the case--i.e. the population standard deviations would have been systematically underestimated and, hence, the Treatment Effect would have been systematically overstated. The important feature noted in this study is not whether use of the proposed variance estimation procedures produce "higher" or "lower" treatment effects. The important feature is that the differences noted between actual and estimated population standard deviations, based on scores for this group of Title I students, were both systematic and substantial. In light of this, and based upon the fact that some local school systems are moving toward utilization of criterion-referenced testing for evaluation of Title I programs, it would appear that a large scale empirical study of the proposed procedures for estimating population standard deviations should be undertaken. In any estimation procedure there will, of course, be errors of estimation. The degree to which those errors are systematic or random should be investigated. Similarly, the magnitude of errors which might be expected, and a system for establishing confidence bands around the estimated standard deviations, should be established. Until such time as these--and other--issues are addressed it would appear that the use of non-normed tests in Title I evaluations will remain a questionable practice. There are very few local Title I project coordinators or evaluators who are willing to take a chance on substantially underestimating the impact of their program, as would have happened for the

program data examined in this study--particularly when the "odds" of making such an underestimate are unknown. It should again be noted that the issues raised in this paper relate only to criterion referenced test data which are to be entered into the uniform evaluation system for aggregation purposes, and do not reflect upon the use of criterion referenced tests for local program evaluation.